Gene 626 (2017) 119–131

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Research paper

Comparative analysis of osteoblast gene expression profiles and Runx2 genomic occupancy of mouse and human osteoblasts in vitro

Kati Tarkkonen[a], Reija Hieta[b], Ville Kytölä[b], Matti Nykter[b,c], Riku Kiviranta[a,d,*]

[a] Institute of Biomedicine, University of Turku, Turku, Finland
[b] GeneVia Technologies, Tampere, Finland
[c] Computational Biology, Institute of Biosciences and Medical Technology (BioMediTech), University of Tampere, Tampere, Finland
[d] Department of Endocrinology, Division of Medicine, Turku University Hospital, Turku, Finland

ARTICLE INFO

Keywords:
Osteoblast
Mesenchymal stem cell
MSC
Runx2
ChIP
ChIP-Seq
Transcriptome sequencing
Microarray
Expression profiling

ABSTRACT

Fast progress of the next generation sequencing (NGS) technology has allowed global transcriptional profiling and genome-wide mapping of transcription factor binding sites in various cellular contexts. However, limited number of replicates and high amount of data processing may weaken the significance of the findings. Comparative analyses of independent data sets acquired in the different laboratories would greatly increase the validity of the data. Runx2 is the key transcription factor regulating osteoblast differentiation and bone formation. We performed a comparative analysis of three published Runx2 data sets of chromatin immunoprecipitation followed by deep sequencing (ChIP-seq) analysis in osteoblasts from mouse and human origin. Moreover, we assessed the similarity of the corresponding transcription data of these studies available online. The ChIP-seq data analysis confirmed general features of Runx2 binding, including location at genic vs intergenic regions and abundant Runx2 binding on promoters of the highly expressed genes. We also found high frequency of Runx2 DNA binding without a consensus Runx2 motif at the binding site. Importantly, mouse and human Runx2 showed moderately similar binding patterns in terms of peak-associated closest genes and their associated genomic ontology (GO) pathways. Accordingly, the gene expression profiles were highly similar and osteoblastic phenotype was prominent in the differentiated stage in both species. In conclusion, ChIP-seq method shows good reproducibility in the context of mature osteoblasts, and mouse and human osteoblast models resemble each other closely in Runx2 binding and in gene expression profiles, supporting the use of these models as adequate tools in studying osteoblast differentiation.

1. Introduction

Fast development and easy availability of high throughput sequencing technologies has allowed rapid accumulation of unbiased genome wide transcription data of several cell and tissue types. Use of chromatin immunoprecipitation combined to NGS sequencing (ChIP-seq) in turn has resulted in rapidly accumulating research literature on global transcription factor (TF) binding site mapping of many different cell types (Cao et al., 2010; Handoko et al., 2011; Heinz and Glass, 2012; Lin et al., 2010; Mikkelsen et al., 2010). Despite the fast progress in the field, there has been relatively few studies utilizing ChIP-seq in cells of the osteogenic lineage, reflecting perhaps the challenging sample material mature bone matrix producing osteoblasts represent. Nevertheless, ChIP-seq analyses of Runx2 (Håkelien et al., 2014; Meyer et al., 2014b; Wu et al., 2014), C/EBPβ (Meyer et al., 2014b), VDR and RXR (Meyer et al., 2014a) in osteoblasts have recently been reported. Because of the high amount of data NGS-approaches produce, it is evident that there are several putative candidate genes and mechanisms to be followed by future studies, eventually generating vast amounts of new information on the regulation of osteoblastogenesis. However, the choice of significant candidates to take forward is challenging as the statistical power remains usually low because generally only few replicate samples are included in the individual experiments. Thus the more studies are published and compared to previous data by bioinformatic tools and approaches, the more consistent and reliable information concerning individual genes and pathways can emerge.

Osteoblasts originate from mesenchymal stem cells (MSC) that can give rise to a number of specialized cell types such as adipocytes,

Abbreviations: NGS, next generation sequencing; ChIP, chromatin immunoprecipitation; ChIP-seq, chromatin immunoprecipitation followed by deep sequencing; TF, transcription factor; MSC, mesenchymal stem cell; DE, differentially expressed; TSS, transcriptional start site; ALP, alkaline phosphatase; shRNA, short hairpin RNA; GREAT, genomic regions enrichment of annotations tool; ENA, european nucleotide archive; GO, genomic ontology; YAP, Yes-associated protein; TAZ, transcriptional coactivator with PDZ-binding motif

* Corresponding author at: Institute of Biomedicine, University of Turku, FI-20520 Turku, Finland.
E-mail address: riku.kiviranta@utu.fi (R. Kiviranta).

http://dx.doi.org/10.1016/j.gene.2017.05.028
Received 27 January 2017; Received in revised form 10 May 2017; Accepted 10 May 2017
Available online 11 May 2017

Table I
Summary of the experiments and data provided by the original articles.

Study	Cell model	ChIP-seq time points	Gene expression data: method, time points	ENA study number
Håkelien et al.: *The Regulatory Landscape of Osteogenic Differentiation*	iMSC#3 Immortalized human mesenchymal stem cell line	28 d	RNA-seq, 0 d cells 28 d cells	ERP003787
Meyer et al.: *The RUNX2 Cistrome in Osteoblasts: characterization, down-regulation following differentiation, and relationship to gene expression*	MC3T3-E1 Mouse preosteoblastic cell line	0 d 15 d	Microarray (Mouse 385 K microarray, Roche Nimblegen) 0 d cells (POB) 15 d cells (OB)	SRP016885
Wu et al.: *Genomic Occupancy of Runx2 with Global Expression Profiling Identifies a Novel Dimension to Control of Osteoblastogenesis*	MC3T3-E1 Mouse preosteoblastic cell line, subclone 4	0 d 9 d 28 d	Microarray (GeneChip Mouse Gene 1.0 ST Array rev.4, Affymetrix) ShRunx2 (9d) cells ShScr (0 d and 9 d) cells	SRP035343

myoblasts and chondrocytes. Numerous hormones, growth factors and cytokines regulate the differentiation of an MSC to a specific cell type, driven by a series of transcription factors that control phenotype-specific gene expression. In case of osteoblasts, early bipotent chondro-osteogenic progenitor cells express transcription factor SOX9, which is then followed by the expression of Runt family transcription factor (Runx2) and its downstream target Osterix (Osx) in preosteoblasts (Long, 2012). Runx2 and Osx are both indispensable for osteoblast differentiation as their null mutant mice show total absence of bone and osteoblasts (Ducy, 2000; Nakashima et al., 2002).

Runx2 belongs to Runt family of transcription factors (Runx1–3) that regulate development and differentiation of many different cell lineages. Runx2 protein contains a conserved 128 amino acid Runt domain, which is responsible for the DNA binding and heterodimerization with CBFβ, that enhances Runx2 binding to DNA (reviewed in Cohen, 2009). Importantly, Runx2 can interact with several proteins including co-regulatory proteins and chromatin remodeling factors, leading to complex role in regulating bone specific genes and differentiation. Control of Runx2 expression is complex, and includes epigenetic mechanisms such as miRNAs and several histone modifying enzymes (Huang et al., 2009; Rojas et al., 2015; Yang et al., 2013). Moreover, Runx2 activity is regulated by several posttranslational modifications such as phosphorylation, acetylation, ubiquitination (reviewed in Jonason et al., 2009) and sumoylation (Kim et al., 2014).

Mouse MC3T3-E1 cell line is a clonal non-transformed cell line established from new born mouse calvaria (Sudo et al., 1983) and is commonly used for studying osteoblast differentiation in vitro. These cells represent Runx2 positive pre-osteoblasts committed to osteogenic lineage. MC3T3-E1 cells maturate to mineralizing osteoblasts in the presence of standard osteogenic medium containing ascorbic acid and Na-β-glycerophosphate. Of human origin, there are few non-transformed osteoblast cell lines. Thus differentiation of human cells is often investigated by using primary human osteoblasts, immortalized human osteoblast lines (HOBs) or immortalized mesenchymal stem cells (iMSCs). Critical evaluation of the current osteoblast cell culture models is important for further development of reliable and adequate in vitro models, not only for high standard basic research but also for use in pharmaceutical and biomaterial research. Typically, phenotypic assessment of osteoblastic cells include measurement of the expression level of osteoblastic genes (Runx2, Sp7, Ocn, Opn), alkaline phosphatase (ALP) activity and formation of mineralized bone matrix in differentiation cultures. Comparison of different models has often been limited to these few phenotypic properties.

Intriguingly, three papers were published reporting ChIP-seq analyses of genome-wide Runx2 binding in the context of osteoblast differentiation in spring 2014. Two of them, papers from the laboratories of Wesley Pike (Meyer et al., 2014b) and Jane Lian (Wu et al., 2014), described Runx2 ChIP-seq analyses of mouse MC3T3-E1 in undifferentiated and differentiated state. Paper of Håkelien and co-workers in turn (Håkelien et al., 2014), described mapping of Runx2 binding sites in iMSCs (Skarn et al., 2014) differentiated to mature osteoblasts in vitro. In general, there was high degree of similarity in the results reported in these papers, but due to the publishing dates close to each other, the data sets were not directly compared by any of the authors. Our approach was to objectively evaluate and compare the published global gene transcription and Runx2 ChIP-seq data sets collected from these three selected osteoblast studies. Our specific aims were 1) to study differences in the Runx2 genomic occupancy in mouse MC3T3-E1 osteoblasts during their differentiation in data sets produced by two different laboratories, to evaluate the reproducibility of ChIP-seq experiments in this model, 2) to examine the interspecies differences in Runx2 binding patterns and target genes between human and mouse samples and 3) to evaluate the similarity of gene expression profiles of undifferentiated and differentiated osteoblasts from mouse and human origin.

2. Methods

2.1. Original data

Summary of the published Runx2 ChIP-seq data and gene expression data used in the study are described in Table I.

2.1.1. Data pre-processing and alignment

Raw ChIP-seq sequencing reads were downloaded from European Nucleotide Archive (ENA) and the reads were subjected to quality control using FastQC software (Andrews, 2010). Alignments to reference genomes were performed using bowtie2 (version 2.1.0) (Langmead and Salzberg, 2012), with reads from Meyer et al. and Wu et al. samples aligned to mouse mm9 genome assembly and reads from Håkelien et al. samples aligned to human hg19 genome assembly. All samples were subjected to PCR duplicated removal, after which FastQC quality control was applied again. Alignment statistics of the Bowtie2 alignments are shown in Supplemental Table I. Obtained alignments were inspected visually using Integrative Genomics Viewer (IGV) (Robinson et al., 2011).

2.1.2. ChIP-seq peak detection and annotation

Peaks were detected from the alignment files by MACS 1.4.2 software (Zhang et al., 2008) using default settings. For Håkelien et al. sample, the peak enrichment was determined relative to a control ChIP with anti-H3 antibody in day 28 iMSC3 cells. For Meyer et al. samples, the peak enrichments were determined relative to a control with IgG antibody in day 0 and day 15 cells. For Wu et al. samples, the peak enrichment was determined relative to sonicated input DNA from day 9 MC3T3-E1 cells. Peaks with p-value $< 10^{-10}$ were used in all downstream analyses.

Peaks were annotated to the closest protein coding genes using Bedtools closest (Quinlan and Hall, 2010) tool and gene annotations of the peaks lists sorted by ascending p-value were compared. Because of

higher deviation of Meyer et al. 15d samples, the replicate with highest FRiP (Fraction of Reads in Peaks) scores and largest amount of high confidence peaks were used for comparisons with other samples. For cross-species sample comparison, human gene symbols were translated to orthologous mouse gene symbols. For mouse samples from Meyer et al. and Wu et al. experiments, the overlap of the MACS detected peak regions was also compared directly using R package ChIPSeeker (Yu et al., 2015), which calculated statistical significance of the peak overlap.

The genomic locations of Runx2 binding peaks were examined using a peak annotation function of HOMER (Hypergeometric Optimization of Motif EnRichment) software (Heinz et al., 2010). Similar analysis was also performed using Genomic Regions Enrichment of Annotations Tool (GREAT) (McLean et al., 2010). The analysis was done using default settings that assign a basal regulatory domain of − 5 kb and + 1 kb of the transcriptional start site (TSS) and extend it in both directions to the nearest gene's basal regulatory domain but no > 1000 kb distance. The ChIP-seq peak regions were then associated with the genes in whose regulatory domains they laid. GREAT was also used to find GO annotations that are enriched among the genes near the ChIP-seq peak regions. For GO enrichment analysis, the assigned gene regulatory domain extended in both directions to the midpoint between the gene's TSS and the nearest gene's TSS but no > 50 kb distance.

2.1.3. Runx2 binding motif scan

Enrichment analysis of conserved Runx2 binding sequence motifs was performed using HOMER software (Heinz et al., 2010). The analysis was done using default settings and repeat-masked sequence.

2.2. DNA microarray data analysis

2.2.1. Processing of Meyer et al. microarray data

Meyer et al. provided gene expression data obtained from mouse 385 k microarrays (Roche NimbleGen) of pre-osteoblast stage MC3T3-E1 cells (POB) and osteoblast stage MC3T3-E1 cells (OB) after 15 days of differentiation. The data consisted of normalized gene expression values with and without log2 transformation as well as log2 fold-change values and was used in this analysis as provided. For some genes there were more than one fold change value in the table provided by Meyer et al., which were apparently obtained from multiple probes detecting the same gene. These values were averaged by taking a geometric mean of the fold change and the resulting single fold change value for each gene was log2-transformed.

2.2.2. Processing of Wu et al. microarray data

Wu et al. provided gene expression data obtained from GeneChip Mouse Gene 1.0 ST Array rev.4 (Affymetrix) as RMA normalized values in the Gene Expression Omnibus (GEO) database (accession number GSE53982). The normalized, log2-transformed gene expression values were used for further analyses. The samples used for this analysis were day 0 MC3T3-E1 cells infected with control Scramble-shRNA and the same cells after 9 days of differentiation. Differential expression between the day 0 and day 9 Scr-shRNA cells was determined using R Bioconductor package limma (Ritchie et al., 2015) with obtained *p*-values adjusted for multiple testing by Benjamini-Hochberg method (Benjamini and Hochberg, 1995). Differential expression was defined as significant if the adjusted *p*-value was below the significance level 0.05 and absolute log2 fold change was at least 1.

2.2.3. Processing of Håkelien et al. RNA-seq data

For Håkelien et al. study, raw RNA-seq read files of day 0 undifferentiated iMSC3 cells and day 28 differentiated iMSC3 cells (two replicates of each) were downloaded from European Nucleotide Archive (ENA) (accession number ERA246830). Quality control of the reads was performed using FastQC software. The quality of all read files was found to be reasonable and no pre-processing was considered necessary. The reads were aligned against human hg19 genome using Bowtie2. Bedtools multicov tool (Quinlan and Hall, 2010) was used for counting the reads aligned to all exonic regions in the genome. The counts from all exons belonging to the same gene were then summarized to get gene-level read counts, and the obtained counts were FPKM (Fragments Per Kilobase of exon per Million fragments mapped) normalized. The non-normalized read counts were used for differential expression analysis by R Bioconductor package DESeq2 (version 1.6.3) (Love et al., 2014). The obtained *p*-values were adjusted for multiple testing by Benjamini-Hochberg method. Differential expression was defined as significant if the adjusted *p*-value was below the significance level 0.05 and absolute log2 fold change was at least 1.

2.3. Gene expression data comparison

For comparison of the gene expression profiles from Meyer et al., Wu et al. and Håkelien et al. studies, the gene symbols in Håkelien et al. data were mapped to orthologous mouse gene symbols. Log2-transformed expression values of all samples, scaled to the same range, were hierarchically clustered based on computed Euclidian distance of the samples with complete linkage method.

Further comparison of the gene expression profiles between the three studies utilizing different platforms and two different species was done using R Bioconductor package OrderedList (Yang et al., 2006) with function that detects similarities between two ordered gene lists. The function compares two-ranked list, in this case gene lists ordered by decreasing gene expression value, and a similarity score is assigned based on the number of overlapping genes in the top ranks. Random scores are computed by comparing one list to the randomly shuffled second list, and based on the random scores an empirical *p*-value can be computed for the observed score. 5000 permutations were used here for estimating the empirical *p*-values.

Pathway enrichment analyses for the three gene expression data sets were performed by GSEAPreranked, which runs Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005) against a user-supplied, ranked list of genes. For the analysis, the gene expression lists of all samples containing all the genes were ordered based on the log2 expression values in descending order and mouse gene symbols in Meyer and Wu gene lists were mapped to orthologous human gene symbols. Enrichment was tested for all available KEGG pathway gene sets with number of permutations set to 1000 and enrichment statistic set as classic. For comparison of the pathway enrichment results, the results were filtered using a stringent false discovery rate (FDR) cutoff of 0.01 and ordered by ascending FDR.

2.4. Comparison of gene expression to Runx2 occupancy at gene promoters

The relationship between Runx2 binding to gene promoter regions and gene expression was studied using the undifferentiated and differentiated mouse samples from Meyer and Wu studies. Gene expression lists ordered by descending log2 expression values were subset into four groups: (1) Genes with expression level within the highest 20% in undifferentiated cells, (2) genes with expression level within the lowest 20% in undifferentiated cells, (3) genes with expression level within the highest 20% in differentiated cells and (4) genes with expression level within the lowest 20% in differentiated cells. The number of Runx2 ChIP-seq peaks on gene promoter areas in samples at different time points (days 0 and 15 from Meyer study and days 0, 9 and 28 from Wu study) were analyzed and the peak distributions between the four gene groups were compared.

3. Results and discussion

3.1. General comparison of the experimental set up between the studies

The experimental setup for osteoblast differentiation culture was

Table II
ChIP-seq peak detection results.

	Meyer et al.				Wu et al.			Håkelien et al.
MACS detection Peak count	D0_1	D0_2	D15_1	D15_2	D0	D9	D28	D28
Cut off $p < 10–5$	38,211	34,296	46,249	34,296	23,667	96,368	49,425	122,325
Cut off $p < 10–10$	26,297	22,966	9810	22,966	14,396	53,887	32,894	80,061
FRiP (Fraction of Reads in Peaks)	17.3%	13.6%	4.3%	11.3%	4.1%	21.0%	13.0%	48.7%

highly similar in all three Runx2 ChIP-seq studies analyzed here, with some variation in the model cell line, in the length of the culture and composition of the osteogenic medium (Table I). In the two MC3T3-E1 studies, different clones of MC3T3-E1 cells were used, Wu et al. using a later isolated rapidly mineralizing subclone from the original MC3T3-E1 cells (Wang et al., 1999). iMSC#3 line used in Håkelien et al. study was produced by immortalizing human bone marrow derived stromal cells by retroviral transduction of telomerase reverse transcriptase (TERT). These cells can differentiate to osteoblasts and adipocytes (Skarn et al., 2014). Osteogenic medium contained ascorbic acid and Na-β-glycerophosphate in all studies, but dexamethasone was used only in the cultures of iMSCs. Osteoblast maturation was verified similarly in all studies by standard methods of ALP, alizarin red and Von Kossa staining to demonstrate lineage commitment and bone nodule formation, respectively.

For Runx2 ChIP-seq, in both MC3T3-E1 studies the first cell stage chosen for analysis were the undifferentiated confluent cells, and the endpoint for sample collection was either 15d (Meyer et al.) or 28d (Wu et al.) after differentiation induction. In addition, Wu et al. included a timepoint representing matrix-producing cells prior to mineralization (9d). In iMSCs, only the 28d differentiated mature osteoblasts were analyzed for Runx2 binding. Two different Runx2 antibodies were used in the studies discussed here, a rabbit polyclonal antibody for MC3T3-E1 cells, and a mouse monoclonal antibody for iMSCs. Notably, the epitope for both these antibodies is in the very same region of Runx2 protein. Despite standardized protocols for some cell types, chromatin immunoprecipitation (ChIP) is a challenging method requiring extensive optimization of sample and antibody-specific conditions including e.g. crosslinking time and temperature, cell lysis, chromatin shearing and washing conditions. Furthermore, special requirements concerning input DNA size and quality have to be met in order to produce good quality sequencing libraries for ChIP-seq. The late-stage osteoblastic cells are tightly surrounded by the collagenous matrix and thus are difficult sample material for ChIP. In the studies discussed here, there was some variability in the ChIP procedures including e.g. cell lysis and sonication conditions, which might have an impact on the sample quality and results, but these aspects are not in the focus in our study.

The method and experimental set up for transcriptional profiling of osteoblastic cells as well as approaches to correlate Runx2 binding site data to transcriptome differed between the studies. Meyer et al. performed microarray expression analysis of MC3T3-E1 at undifferentiated and differentiated state while Håkelien et al. used RNA-seq analysis of iMSC#3 cells of undifferentiated and 28d differentiated cells. Wu et al. in turn did a microarray gene expression profiling of 9d differentiated MC3T3-E1 cells with shRNA mediated knockdown of Runx2, and compared the changes in gene expression to control shRNA cells. In order to compare gene expression profiles between the studies, only the control scrambled shRNA data of Wu et al. study was analyzed here, corresponding to the parental MC3T3-E1 cells of Meyer et al. study.

3.2. ChIP-seq analysis

3.2.1. Quality of the sequencing data

Raw ChIP-seq sequencing reads were downloaded from ENA and subjected to quality control. In summary, Wu et al. and Håkelien et al. samples showed very high quality throughout the reads. Meyer et al. samples had noticeable decrease in quality towards the end of reads, especially in 15d replicate sample 1, but as the majority of bases in reads were of very good quality and the median quality scores were within the very good quality range, no further pre-processing was considered necessary.

3.2.2. Comparison of peak detection results

After duplicate removal the aligned reads were used for peak detection using MACS (Model-Based Analysis of ChIP-seq) (Zhang et al., 2008). Peak detection results for all samples are shown in Table II. In summary, large amounts of Runx2 peaks were detected in all samples even if more stringent p-value cutoff of 10^{-10} was used. Noticeable is the reduction in peak count in Meyer day 15 replicate 1 sample after more stringent cutoff was applied, in accordance with the lower quality of reads that was observed in quality control. A simple quality metrics FrIP for ChIP-seq data is also shown in Table II. The recommendation for this metrics is $\geq 1\%$ by ENCODE consortium for a successful experiment (Landt et al., 2012), and the scores of all samples exceed this limit. Peaks with p-value $< 10^{-10}$ were used in all downstream analyses.

In the published paper Meyer et al. reported contraction of the Runx2 cistrome during differentiation, demonstrated by 12,674 and 6272 genomic regions with Runx2 binding in the early and late state osteoblasts, respectively (Meyer et al., 2014b). Wu et al. in turn reported approximately 25,000 significantly enriched regions at day 0 and an increase in the number of Runx2 bound sites in later timepoints, yielding up to 80,000 Runx2 enriched regions in MC3T3-E1 cells (Wu et al., 2014). In the re-analysis performed here, we did not observe as high variation in the detected peak numbers between the two MC3T3-E1 studies as reported previously, suggesting the difference in the peak numbers between the original studies to be related to the data analysis rather than to a major difference in Runx2 binding abundance (Table II). In our analysis, Wu et al. samples showed increased number of Runx2 peaks in D9 and D28 samples compared to D0 sample, whereas peak number in Meyer et al. study stayed constant between the timepoints (when excluding the one D15 replicate of lower quality) (Table II). In the original papers, there was a minor difference in Runx2 protein expression pattern, the protein level being constant in Meyer's paper but showing increase in Wu's paper, which is in accordance with the dynamics in Runx2 peak numbers observed. In iMSC#3 cells, we found large number of Runx2 peaks (80061) and very high FRiP value (48,7%), suggesting a successful ChIP experiment. However, Håkelien et al. reported only 9549 peaks in the 50 kb region of closest 5′ gene end, suggesting major differences in the analysis methods and/or filtering parameters used. Thus, we will not further compare our iMSC analysis results to the published results, but use our de novo analysis results to the analysis of mouse MC3T3-E1 cells in order to perform inter-species comparisons.

3.2.3. Comparison of annotated genes close to Runx2 peaks

Peaks were annotated to the closest protein coding genes using bedtools closest tool (Quinlan and Hall, 2010). Annotated peak lists were sorted by ascending p-value so that the most enriched peaks of

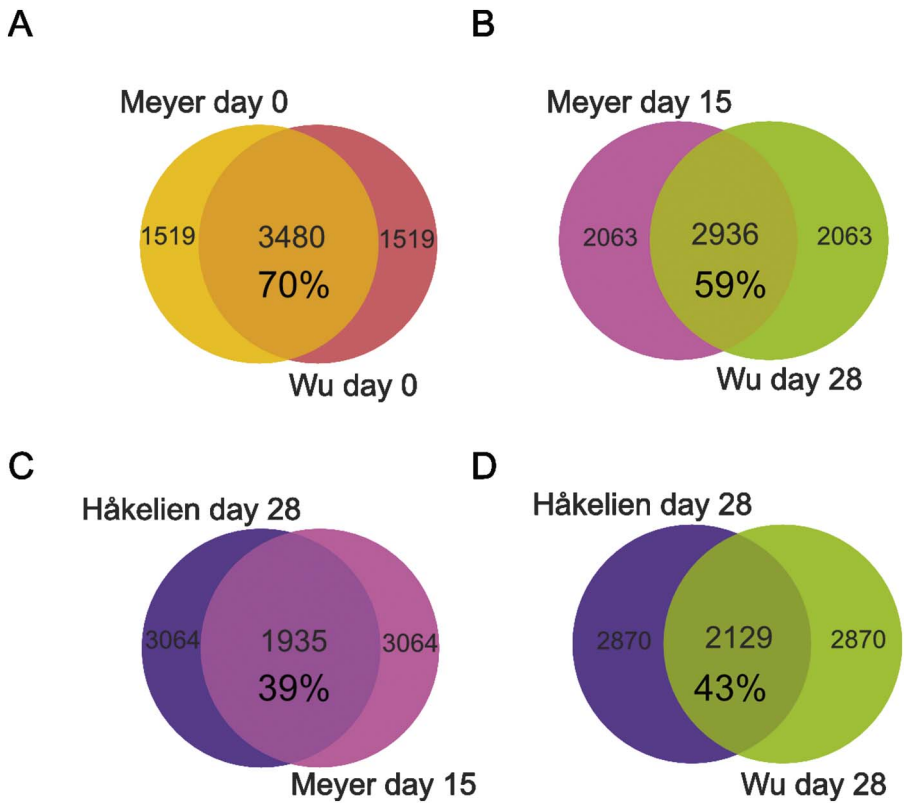

Fig. 1. Venn diagrams of peak annotation comparisons. Gene annotations of the top 5000 most enriched Runx2 ChIP-seq peaks were used for comparisons: A–B) Undifferentiated and differentiated MC3T3-E1 cells of Meyer et al. and Wu et al. Samples were compared, and C–D) Differentiated Håkelien et al. iMSC3 cells were compared to Meyer et al. and Wu et al. differentiated MC3T3-E1 cells.

each sample could be compared. To allow cross-species comparisons, the human gene symbols were translated to orthologous mouse gene symbols. Venn diagrams of the unique and common genes annotated to the top 5000 most enriched peaks are shown in Fig. 1. Meyer et al. and Wu et al. undifferentiated mouse samples seem to be fairly similar, with 70% of top 5000 peak annotations shared between the samples (Fig. 1A), whereas differentiated samples (Meyer et al. day 15 and Wu et al. day 28) show a smaller, 59%, overlap (Fig. 1B). Comparison of Meyer et al. undifferentiated and differentiated samples shows an overlap of 74% in annotations, with corresponding comparisons between Wu samples showing 68% and 71% overlaps for day 9 and day 28 samples, respectively (Supplementary Fig. 1). In conclusion, the comparisons between the differentiation stages suggest that Runx2 constitutively occupies more than half of the total Runx2 bound genes regardless of the differentiation stage.

Comparison of peak annotations of human differentiated iMSC3 samples to differentiated mouse MC3T3-E1 samples from both Meyer and Wu shows overlaps of 39% and 43% in top 5000 annotated genes, respectively (Fig. 1C,D). The overlap percentages are, as could be expected, lower than those of within-species comparisons and in line with previous findings that have shown that most TF binding events are species-specific (Wilson and Odom, 2009). However, altogether 1344 genes from top 5000 annotated genes were in common for all samples in the last differentiation time point, suggesting substantial similarity in Runx2 peak associated genes between the species and experiments. This could suggest that these genes are especially important for the Runx2-mediated regulation of osteoblast differentiation.

For Meyer and Wu mouse samples it was possible to compare directly the overlap of the peak regions. The results indicate that peaks in undifferentiated and differentiated mouse samples have a highly significant overlap both within and between the studies (Table III). For example, when comparing undifferentiated cells to differentiated cells in each study, 88% (Wu et al. study) and 67% (Meyer et al. study) of the peak regions at day 0 were present also in differentiated cells. Moreover, at day 0, 88% of peaks observed in Wu et al. study are detected also in Meyer et al. day 0 sample. In the differentiated cells, corresponding overlap was 50%, possibly related to the variable time course and the different MC3T3-E1 cell clones used in the two studies as well as the challenging sample material at the late time points. Importantly, these results indicate that ChIP-seq experiments in MC3T3-E1 cells show relatively high reproducibility in two individual studies and again that majority of the Runx2 bound sites are constant during differentiation.

Table III
Statistical test of ChIP-seq peak overlap in mouse samples.

Query sample	Target sample	Query peak count	Target peak count	Overlapping peaks	*p*-value	Adjusted *p*-value
Meyer DO_1	Meyer D15_2	26,297	34,871	17,569	0	0
Wu D0	Wu D9	14,396	53,887	13,255	0	0
Wu D0	Wu D28	14,396	32,894	12,714	0	0
Meyer D0_1	Wu D0	26,297	14,396	12,609	0	0
Meyer D15-_2	Wu D28	34,871	32,894	16,577	0	0

3.2.4. Comparison of genomic locations of Runx2 peaks

The genomic locations of Runx2 binding peaks were examined using a peak annotation function of HOMER software. The distributions of Runx2 peaks are displayed in Fig. 2A. In accordance with the original articles, the vast majority of Runx2 binding occurred at intergenic and intronic regions. The greatest variation between the differentiation stages was in Runx2 occupancy at promoters, with higher occupancy in

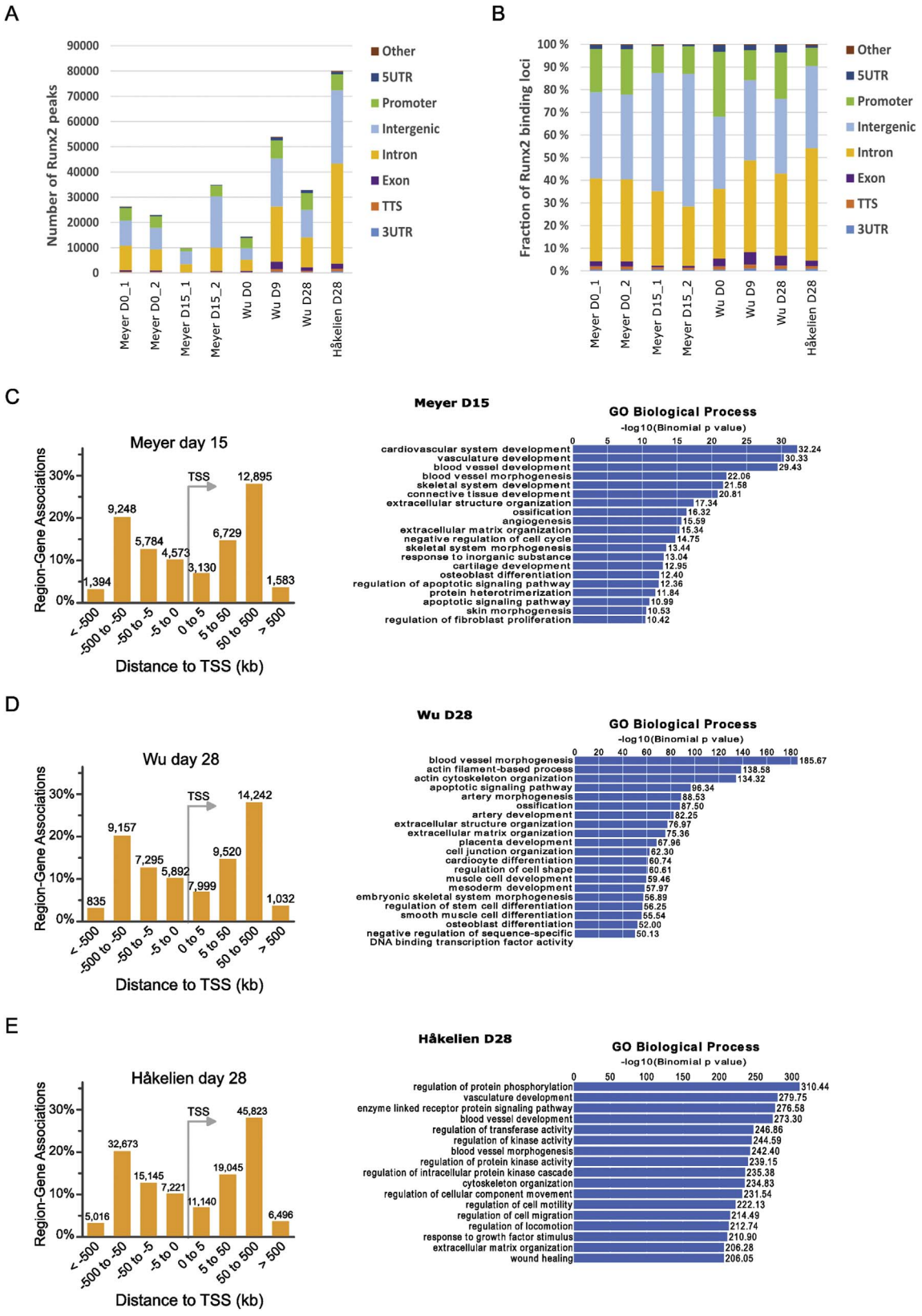

Fig. 2. Genome-wide profile of Runx2 occupancy. A–B) Distribution of Runx2 binding peaks across mouse and human genomes were classified into eight categories: Exon, Intron, Promoter (− 1 kb to + 100 bp from transcription start site), TTS region (− 100 bp to + 1 kb from transcription termination site), 5′ UTR exon, 3′ UTR exon, Intergenic region and other regions. Peak distributions in all samples are plotted by peak number (A) and by the percentage of total peaks (B). C–E) Analysis of the Runx2 peak association to genomic regions using GREAT at the last time point in each of the three studies. In the analysis, a basal gene regulatory domain of − 5 kb and 1 kb of the TSS was assigned and extended in both directions to the nearest gene's basal regulatory domain with a maximum distance of 1000 kb. Runx2 peak regions where then associated with the genes in whose regulatory domains they are located. In addition, GO Biological Process term enrichment analysis of the last time point samples in each study using GREAT is presented (right panels). In this analysis, the assigned gene regulatory domain extended in both directions to the midpoint between the gene's TSS and the nearest gene's TSS with a maximum distance of 50 kb. The figure shows the most enriched GO terms among the genes near the ChIP-seq peak regions.

Table IV
Genome-wide Runx2 peak distributions by the percentage of total peaks.

	Meyer				Wu			Håkelien
	D0_1	D0_2	D15_1	D15_2	D0	D9	D28	D28
3′UTR	0.6%	0.6%	0.6%	0.5%	0.5%	1.0%	0.8%	0.8%
TTS	1.4%	1.4%	1.0%	0.9%	1.5%	1.8%	1.5%	1.2%
Exon	2.2%	2.2%	0.8%	0.9%	3.4%	5.5%	4.3%	2.5%
Intron	36.5%	36.3%	32.9%	26.2%	30.7%	40.5%	36.2%	49.6%
Intergenic	38.1%	37.5%	52.2%	58.6%	32.0%	35.4%	33.1%	36.3%
Promoter	19.1%	19.9%	11.8%	12.1%	28.6%	13.3%	20.4%	8.0%
5′UTR	1.7%	1.9%	0.5%	0.7%	3.0%	2.3%	3.3%	1.1%
Other	0.3%	0.3%	0.3%	0.2%	0.3%	0.3%	0.4%	0.5%
Total	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

undifferentiated cells in both MC3T3-E1 studies (Fig. 2A,B, Table IV). Similar analysis was performed using GREAT software, which finds GO annotations that are enriched among the genes near the ChIP-seq peak regions. Graphs showing the distance between peak regions and their putatively regulated genes as well as the most enriched biological process GO terms obtained from GREAT analysis for differentiated cells are shown in Fig. 2C–E. In summary, the pattern of Runx2 peak enrichment was very similar between the experiments, showing clear enrichment of Runx2 binding in the vicinity of TSS. Among enriched GO terms in biological processes category there were both common and closely related terms and some unique pathways, reflecting the overlap of the annotated genes close to the peaks (Fig. 1). The enriched terms such as RNA related processes, cell- and tissue morphology related pathways, resembled closely the published data of each report. The minor differences in the term enrichments to the original articles results might be due to differences in performing the analysis, for example Wu et al. used time-dependent dynamic clusters of Runx2 peaks, whereas in our analysis all Runx2 peaks at certain time point were included.

3.2.5. *Runx2 binding motif enrichment in Runx2 peak regions*

ChIP-seq peaks were scanned for conserved TF sequence motifs using HOMER software. HOMER includes a motif database that is mostly based on the analysis of public ChIP-Seq data sets (Heinz et al., 2010). The full tables of the results of the motif scan are presented upon request. In mouse samples, the most significantly enriched motifs were three motifs that have been experimentally associated with binding of RUNX-family of transcription factors ((A/T/C)TGTGGTT(A/T); (G/T)(T/C)TGTGGTTT; CTGTGGTTT(G/C))), presented in Table V. These motifs were present in altogether 36–54% of Runx2 peaks in all MC3T3-E1 samples (Table VI). In the motif scan, two mismatches to the consensus sequence were allowed in the analysis, leading to inclusion of high number of different variant motifs as Runx2 motifs, referred hereafter altogether as Runx2 motif. Accordingly, significant enrichment of a classic Runx2 binding core motif TGTGGT at the Runx2-bound regions was reported in both MC3T3-E1 studies previously. In iMSCs on the other hand, the known Runx2 motifs were found in only 14,4% of the peaks and the motifs ranked as most significantly enriched were for AP1 and ETS family transcription factors. Thus, in iMSC sample, although having the highest number of peaks left after quality filtering, only minority of these peaks appear to contain Runx2 binding motifs. This may indicate that the Håkelien et al. sample contains more false positives and would require more stringent filtering of the detected peaks, or that the Runx2 motif in human DNA is highly variable and was not recognized in our analysis.

Table V
Runx2 motifs enriched in HOMER motif enrichment analysis* of MC3T3-E1 cells.

Motif name	Consensus sequence
RUNX2(Runt)/PCa-RUNX2-ChIP-Seq(GSE33889)/ Homer	A/T/C)TGTGGTT(A/T)
RUNX1(Runt)/Jurkat-RUNX1-ChIP-Seq(GSE29180)/ Homer	(G/T)(T/C)TGTGGTTT
RUNX(Runt)/HPC7-Runx1-ChIP-Seq(GSE22178)/ Homer	CTGTGGTTT(G/C)

*The analysis allowed for two mismatch nucleotides in the consensus sequence.

Table VI
Percentages of MACS peak sequences with Runx-related motifs detected by HOMER analysis.

Sample	Total number of peaks	Peaks without RUNX motif	Peaks with RUNX motif	
			Number	Percentage
Meyer D0_1	26,297	16,792	9505	36.1
Meyer D0_2	22,962	14,304	8658	37.7
Meyer D15_1	9810	4514	5296	54.0
Meyer D15_2	34,871	20,813	14,058	40.3
Wu D0	14,396	9219	5177	36.0
Wu D9	53,887	31,310	22,577	41.9
Wu D28	32,894	20,088	12,806	38.9
Håkelien D28	80,061	68,603	11,458	14.3

Interestingly, the motif scan revealed also other significantly enriched TF consensus motifs at Runx2 peak regions. Secondary binding motifs may suggest for example protein-protein interactions, by which Runx2 is recruited to sites without Runx2 binding motif, or possible co-operative binding and/or collaboration of Runx2 with other transcription factors at the same region. When comparing the top 20 enriched motifs in each sample, there were 9 common motifs found in all samples (Table VII). Most of these were related to AP-1 family proteins including Atf3, Fra1 and Jun-AP1 motif. These consensus motifs are found at high frequency in the genome and are involved in a wide variety of cellular processes.

Interestingly, another significantly enriched motif in every data set was for TEAD family factors and especially for TEAD4, present in 9–17% of Runx2 peaks in mouse and in 10% of peaks of human samples. In Håkelien et al. original article, TEAD2 motif enrichment was reported in regions enriched for promoter and enhancer related histone modifications H3K4me3, H3K9ac and H3K27ac at the end of differentiation, and its knockdown was then further shown to cause impaired mineralization in vitro.

TEAD family of transcription factors are effectors in the Hippo signaling pathway regulating organ size by controlling cell proliferation and apoptosis (Landin-Malt et al., 2016). TEADs form a complex with Yes-associated protein (YAP) or its paralog transcriptional co-activator with PDZ binding motif (TAZ) to activate target gene expression. Both YAP and TAZ in turn interact also with Runx2 (Hong et al., 2005; Zaidi et al., 2004) and they have been implicated in regulating MSC differentiation (Hong et al., 2005) and in Wnt and BMP signaling (Varelas, 2014). Very recently, stabilization of TAZ-Runx2 complex was shown to play an important role in promoting osteoblastogenesis (Matsumoto et al., 2016). When inspecting Runx2 peaks with the

Table VII
Top 20 enriched TF motifs in all Runx2 ChIP-seq samples.

Meyer		Wu			Håkelien
0 day	15 day	0 day	9 day	28 day	28 day
RUNX	RUNX	RUNX	RUNX	RUNX	Atf3
RUNX1	RUNX1	RUNX1	RUNX1	RUNX1	BATF
RUNX2	RUNX2	RUNX2	RUNX2	RUNX2	Fra1
RUNX-AML	RUNX-AML	RUNX-AML	RUNX-AML	RUNX-AML	AP-1
TEAD4	TEAD4	Fra1	Fra1	Fra1	Fosl2
TEAD	TEAD	BATF	Atf3	BATF	Jun-AP1
Fra1	TEAD2	Atf3	CTCF	Atf3	Bach2
Fosl2	Atf3	Fosl2	BATF	Fosl2	NF-E2
Jun-AP1	Fra1	Jun-AP1	Fosl2	AP-1	Bach1
BATF	BATF	AP-1	AP-1	Jun-AP1	Nrf2
TEAD2	AP-1	REST-NRSF	Jun-AP1	CTCF	MafK
Atf3	Fosl2	TEAD4	BORIS	Bach2	ERG
AP-1	Jun-AP1	Bach2	Bach2	REST-NRSF	ETV1
CTCF	CTCF	TEAD	NF1	TEAD4	ETS1
Bach2	Ap4	TEAD2	TEAD4	TEAD	TEAD4
NF1-halfsite	BORIS	Fli1	Klf4	Fli1	Fli1
BORIS	Bach2	Elk4	KLF5	BORIS	GABPA
Elk1	Atoh1	Elk1	EKLF	TEAD2	TEAD
ETS	Tcf12	ELF1	TEAD	Ap4	MafA
Elk4	MyoG	ETS	Tlx	Elk1	TEAD2

TEAD4 motif more closely, approximately half of the TEAD4 motif containing peaks contained also Runx2 motif in mouse samples, comprising 4–10% of all Runx2 peaks, whereas in human iMSCs only 1,5% of total Runx2 peaks contained both motifs (Supplemental Table II). These results suggest that TEAD4 and Runx2 interaction might take place at least in part of the Runx2 occupied regions through a common protein complex (TEAD4 motif only regions) or through collaborative binding via putative nearby DNA binding motifs (co-presence of TEAD4 and Runx2 motifs). Given the emerging literature of the importance of TEAD/TAZ/YAP proteins in osteogenesis (Matsumoto et al., 2016; Tang et al., 2016), role of TEAD-Runx2 interaction and protein complex in osteoblasts would be highly interesting target for future studies.

3.2.6. Runx2 occupancy at the promoter regions

Runx2 binding sites where inspected in more detail at promoter regions, which are more likely to be conserved across species than the distal enhancer sites (Cheng et al., 2014). The promoter sequences (2 kb upstream from TSS) of all orthologous pairs of human and mouse genes were retrieved and annotated for the nearest ChIP-seq peaks in mouse and human samples. The results of the extensive analysis are provided in Supplemental Table III. The promoter coordinates, nearest peak region's coordinates, and the distance of the peak from the promoter are shown for all genes together with information indicating the number of samples, in which Runx2 peaks were found. In addition, the number of consensus Runx2 motifs found at each peak region is reported in the table. If the peak distance from the promoter region is annotated as 0, it means that the peak is overlapping with the promoter. Gene expression data from the studies (described in detail in Section 3.3.) have been integrated into the table as well. Another view of the data is presented in Supplemental Table IV, which shows the presence of Runx2 at gene promoters at different time points together with the related gene expression data, offering possibility to assess the dynamics of Runx2 binding at individual gene promoters between the studies. These tables with annotated gene information can be used as a resource, when exploring possible Runx2 binding in the vicinity of any gene of interest. For example, by filtering and organizing columns in Tables SIII and SIV, various groups of genes of interest can be selected based on their Runx2 binding properties and expression profiles.

We next searched for genes that had Runx2 peaks near the promoter regions in all 7 mouse samples indicating high confidence of constitutive Runx2 binding during differentiation using the Supplemental Table IV. These genes and their promoters had several common features. First, most of these Runx2-occupied genes were relatively highly expressed but only minority (11%) of them contained Runx2 motifs in the peak regions. Similarly, the corresponding human orthologues of these genes had nearby Runx2 peaks, were expressed and showed low frequency of Runx2 consensus motifs. These observations suggest either Runx2 motif is highly variant on these sites, or that Runx2 may be recruited to many regions by other mechanisms than by direct binding to Runx2 motifs in the DNA. Whether Runx2 binding is required for the transcription of these genes is not clear. Interestingly, in the study of Wu et al. only 159 genes responded to Runx2 knockdown during differentiation, which is far less than genes occupied by Runx2. As Runx2 has also genome-organizing capabilities (Lian et al., 2003), it may well be that on these other genes Runx2 modifies the genomic landscape rather than directly regulates gene expression.

Nevertheless, Wu et al. reported more Runx2 peaks in the shRunx2 downregulated genes (thus Runx2 upregulated genes) compared to shRunx2 upregulated or nonresponsive genes. Accordingly, when we inspected the abundance of Runx2 peaks and Runx2 motifs close to the shRunx2 responsive/unresponsive genes (Table VIII), the shRunx2 downregulated genes contained indeed more Runx2 peaks per gene, and especially more Runx2 motif containing peaks per gene (average 3,1/gene) than the upregulated or nonresponsive genes (average 1,1/gene), supporting the high frequency of Runx2 motifs together with Runx2 binding to be an strong indicator of direct transcriptional activation by Runx2.

3.3. Gene expression data

The gene expression analysis was performed by using different platforms and experimental setups as described in Section 3.1., resulting in data provided in different formats in on-line repositories. For the comparisons, all gene expression data sets were processed to produce normalized log2-transformed expression values for the expressed genes.

3.3.1. Differential gene expression analysis

Meyer et al. provided a ready-to-use table of gene expression profiles obtained from mouse DNA microarray analysis of pre-osteoblast stage MC3T3-E1 cells (POB) and osteoblast stage MC3T3-E1 cells (OB) after 15 days of differentiation. On the list, there were 721 transcripts including 498 significantly differentially expressed genes in the Meyer et al. study between the differentiation states. Wu et al. in turn performed expression profiling at 0 and 9 days of differentiation of MC3T3-E1 treated with either control scrambled shRNA or Runx2 targeting shRNAs leading to Runx2 downregulation. In the current study, in order to compare gene expression profiles of MC3T3-E1 cells

Table VIII
Number of Runx2 peaks and Runx2 consensus motifs at the Runx2 peaks at the promoter regions of shRunx2 responsive genes.

			Average	200 bp region around peak center		500 bp region around peak center	
shRunx2 response	Gene count	Annotated peaks	Number of peaks/gene	Peaks with Runx motif (s)/gene	% of peaks with Runx motif(s)	Peaks with Runx motif (s)/gene	% of peaks with Runx motif(s)
Down-regulated	44	288	6.5	3.1	47	4.3	65
Up-regulated	115	296	2.6	1.1	44	1.6	61
Non-responsive	20,788	58,886	2.8	1.1	40	1.6	57

Table IX
Top 10 significantly differentially expressed genes in each study.

Meyer et al. 15d vs 0d		Wu et al. 9d vs 0d shScramble		Håkelien et al. 28d vs 0d	
Gene name	Log2FC	Gene name	Log2FC	Gene name	Log2FC
Cd200	6.233142	Lum	6.024314816	MMP13	11.50375
Mmp13	5.055013	Apod	4.675722274	COL10A1	10.72343
Itgbl1	4.405752	F13a1	3.826338637	DPT	10.20353
Lect1	4.374315	Ibsp	3.764573159	OMD	10.10910
Col15a1	4.164052	Mme	3.375728736	CHI3L1	9.27727
Igf2	4.03787	Omd	3.355162249	FAM20A	9.23509
Akap12	3.937885	Pgm5	3.349156029	BRINP1	9.21489
F13a1	3.900387	Ednra	3.213000637	SERPINF1	9.19554
Bmper	− 4.09268	Ppbp	-3.205871844	GGT5	9.07341
Cxcl7	− 4.58329	Anxa8	-3.106959652	MMP7	9.02634

between Meyer and Wu studies, we analyzed only the data of control shRNA expressing cells of Wu study and compared them to the parental MC3T3-E1 cells used in Meyers study. Wu et al. provided normalized signal values and probe IDs. Our analysis of differentially expressed genes filtered by adjusted *p*-value and the 2-fold change yielded 345 differentially expressed (DE) genes between the time points. For Håkelien et al. study, raw RNA-seq read files of day 0 undifferentiated iMSC3 cells and day 28 differentiated iMSC3 cells were available and were analyzed as described in materials and methods. The analysis results were filtered for adjusted *p*-value and the extent of differential expression with 2-fold change. The filtered gene list contained 3039 genes, 1504 up-regulated and 1535 down-regulated, the numbers being in line with the 3157 differentially expressed genes reported in Håkelien et al. study. Top ten differentially expressed genes in each study are shown in Table IX. All DE gene lists are provided as Supplemental Tables V–VII.

3.3.2. Heatmap of gene expression profiles

In order to compare gene expression profiles of all samples in Meyer et al., Wu et al. and Håkelien et al. studies, the gene symbols in Håkelien et al. data were translated to orthologous mouse gene symbols. Log2-transformed expression values of all samples, scaled to the same range and hierarchically clustered, are visualized in a heat map in Fig. 3. The heat map clearly shows that samples coming from the same study are most similar to each other based on their gene expression profiles, regardless of the differentiation stage of the cells. Thus, the effects caused by differences in the cell clones or other technical aspects between the studies mask the effect of differentiation on the overall gene expression profiles even in the samples from the same species. The finding of only moderate changes in gene expression profiles between the differentiation states might also reflect the fact that MC3T3-E1 cells are already committed to the osteogenic lineage and the most striking changes in gene expression from the stem cell stage to preosteoblasts have already passed. Nevertheless, these results demonstrate that the widely distributed cell lines may exhibit clonal shift resulting in phenotypic/epigenetic/transcriptomic changes over time, and highlight the importance of critical evaluation of results obtained from a single cell line.

3.3.3. Similarity of gene expression profiles

In order to assess the similarity of the gene expression profiles between the samples from different studies using different platforms and even from different species, R Bioconductor package OrderedList (Yang et al., 2006) was utilized for detecting similarities of two ordered gene lists. In short, two ranked lists (according to the mRNA expression level) were compared and a similarity score was assigned based on the number of overlapping genes in the top ranks. Plots of overlapping genes in Meyer et al. and Wu et al. day 0 top or bottom ranked gene expression lists are shown in Fig. 4A–B, where the overlap size increases if the gene in top ranks of list 1 is found within the top ranks of list 2. The observed overlap was compared to the expected overlap derived

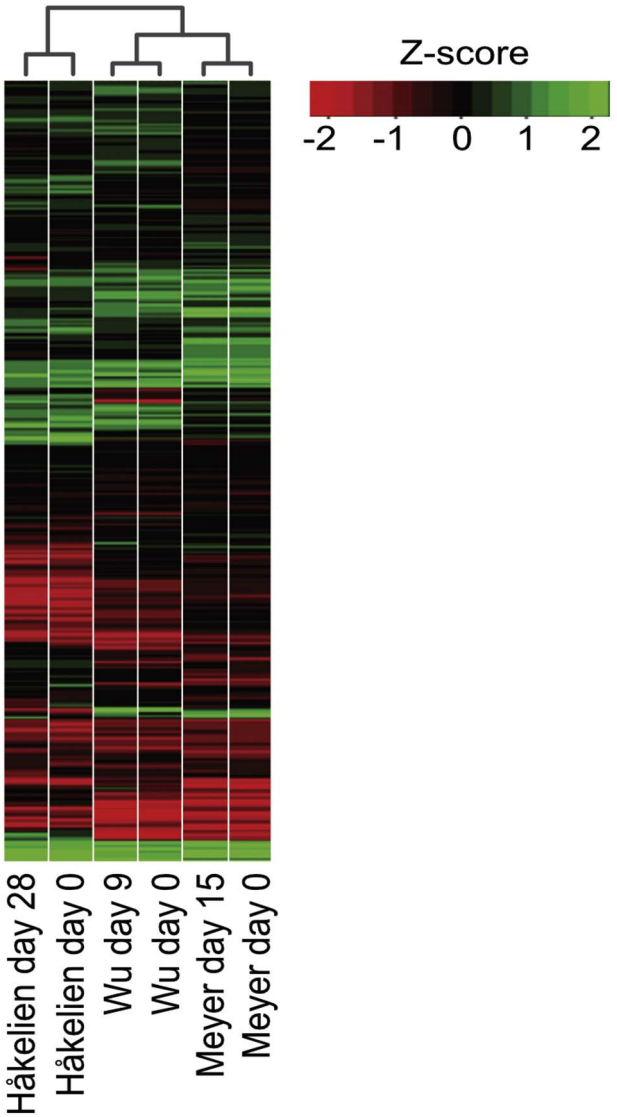

Fig. 3. Heatmap of log2-transformed, normalized gene expression values of Håkelien et al., Wu et al., and Meyer et al. samples. Dendrogram on top of the figure shows the result of hierarchical clustering of the samples using Euclidian distances with complete linkage method.

from a hypergeometric distribution. The significance of the similarity of the lists within top 1000 genes was very high (Fig. 4B), with *p*-value of 0. Within top 100 genes the similarity was lower but still highly significant (Fig. 4A), with *p*-value 0,0026. Number of common genes in these top 100 and top 1000 expressed gene lists was 15 and 271, respectively. Similar comparisons for undifferentiated human osteo-

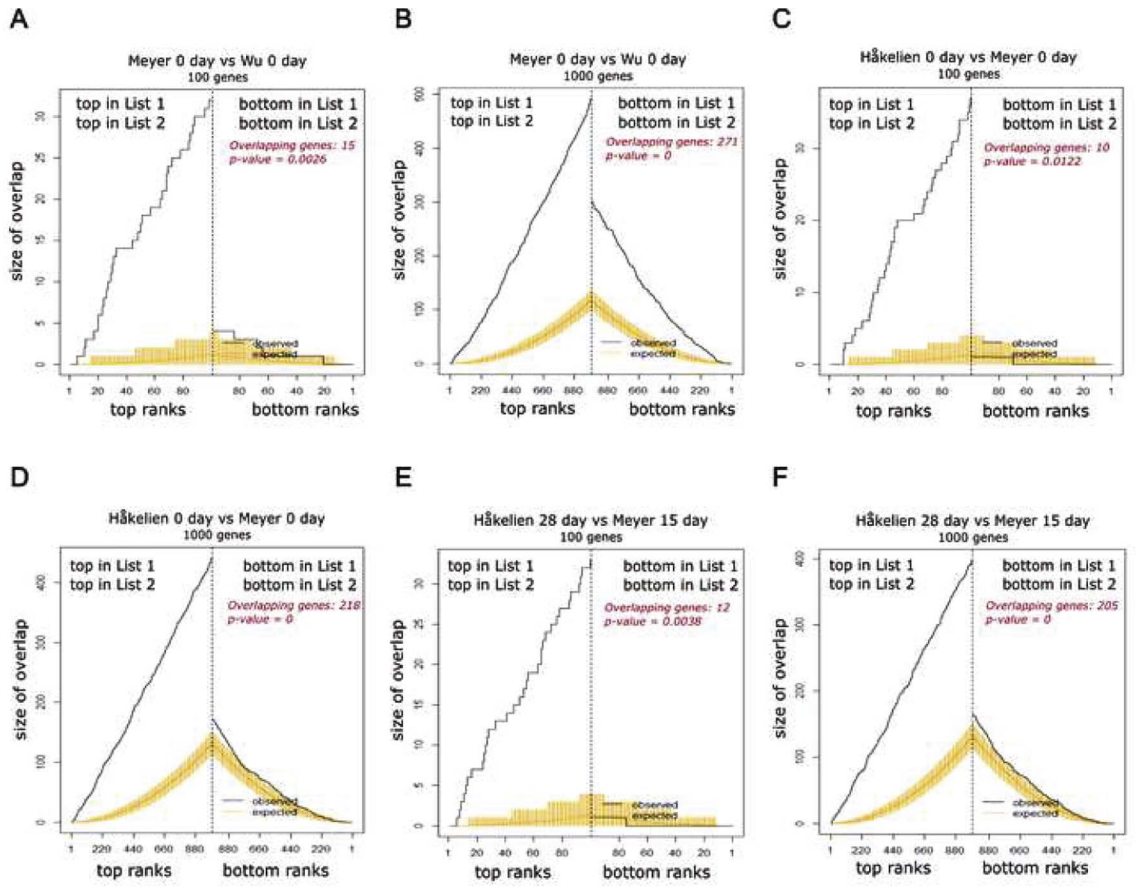

Fig. 4. Comparison of gene expression data of A–B) Meyer et al. day 0 and Wu et al. day 0 samples, C–D) Håkelien et al. day 0 and Meyer et al. day 0 samples and E–F) Håkelien et al. day 28 and Meyer et al. day 15 samples using gene lists ranked by gene expression level in descending order. Overlap of top and bottom 100 and 1000 genes are shown for each comparison as indicated in the figure. The overlap size is drawn as a step function over the respective ranks (black lines), with top ranks corresponding to the genes with highest expression and bottom ranks to genes with lowest expression. The expected overlap and 95% confidence intervals derived empirically from a hypergeometric distribution are shown in orange color. The numbers of overlapping genes in the two lists are displayed as well as the permutation test *p*-values. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

blasts (Håkelien et al. day 0) and undifferentiated mouse osteoblasts (Meyer et al. day 0) and for differentiated human osteoblasts (Håkelien et al. day 28) and differentiated mouse osteoblasts (Meyer et al. day 15) are shown in Fig. 4C–F. In summary, all gene expression lists appeared highly similar within the highly expressed, top-ranked genes. The overlaps of the bottom ranked genes were close to random overlap, which was expected as the genes at the bottom of the ranked lists are mostly very lowly or not at all expressed and appear as random noise.

The similarity of the gene expression profiles was also evaluated based on the pathway enrichment analysis of the gene expression profiles. For the analysis, the gene expression lists of all samples containing all the genes were ordered based on the log2 expression values in descending order, and mouse gene symbols in Meyer et al. and Wu et al. gene lists were translated to orthologous human gene symbols. Fig. 5 shows Venn diagrams of the top 30 most significantly enriched pathways of all samples. The overlaps between the two mouse samples are substantial, with 71% of the pathways in undifferentiated Meyer et al. and Wu et al. samples overlapping (Fig. 5A) and 58% of the pathways in differentiated samples from the same studies overlapping (Fig. 5B). The smaller overlap at the differentiated stage may be due to difference in the time points, as in Meyer et al. cells have been differentiated for 15 days but only 9 days in Wu et al. The comparisons of human iMSC cells with MC3T3-E1 mouse cells show slightly less overlap in the enriched pathways (Fig. 5C–F). Interestingly, for undifferentiated samples from Meyer et al. and Håkelien et al. samples, the overlap is relatively high, 62%, suggesting the phenotype of the cells to resemble each other closely although MC3T3-E1 represent committed pre-osteoblasts and iMSCs multipotent stem cells. Comparison of differentiated human osteoblasts to differentiated mouse osteoblasts from Meyer's and Wu's studies show also a consistent overlap of 58% in the 30 most significantly enriched pathways in both comparisons. However, when evaluating the similarity of gene expression profiles it is important to keep in mind that transcriptomes of different tissues consist of a large amount of ubiquitously expressed genes. For example, in a RNA-seq analysis of multiple tissues, of the 11,000–13,000 expressed genes > 40% were shown to be expressed in all tissue and cell types, and moreover that their mRNAs represented 75% of the total mRNA quantity expressed in any given cell type, suggesting cell type specific mRNAs to account only the minority in a mRNA pool (Ramsköld et al., 2009).

3.3.4. Correlation of Runx2 promoter occupancy and gene expression in MC3T3-E1 cells

Mouse samples from Meyer et al. and Wu et al. studies were then compared for Runx2 binding to promoter regions of genes that are expressed at high or low levels in undifferentiated and differentiated cells. For this purpose, gene expression lists were ordered by decreasing order of log2 expression level. We then analyzed Runx2 occupancy at promoters (− 2 − 0 kb from TSS) of 20% of the highest and lowest expressed genes at each time point. Bar graphs showing the number of unique and common Runx2 occupied genes in the 20% highest/lowest expressed gene groups between the timepoints in each study are shown in Fig. 6. Consistently in both studies, Runx2 binding was remarkably more abundant in promoters of highly expressed gene than of genes

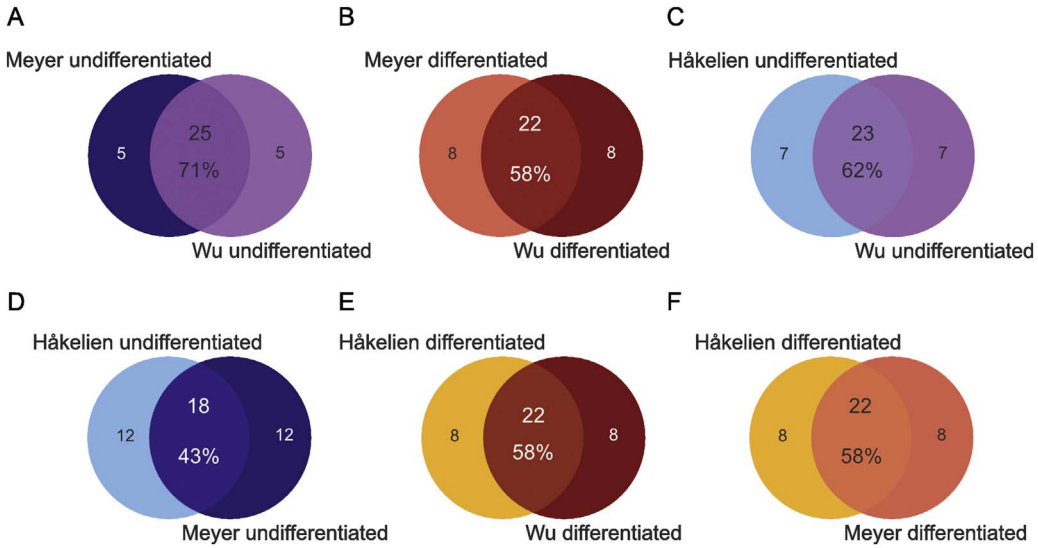

Fig. 5. Venn diagrams of the top 30 most significantly enriched KEGG pathways in samples based on GSEA of ranked gene expression lists. A–B) The mouse osteoblast samples at both differentiation stages from Meyer et al. and Wu et al. studies are compared. C–F) The human osteoblast samples from Håkelien et al. study are compared to mouse osteoblast samples from Wu et al. and Meyer et al. studies at both differentiation stages. Percentage of overlapped pathways is displayed on the intersections.

with low expression. The pattern of Runx2 promoter peak distribution was almost identical in undifferentiated and differentiated cells, which is in accordance with the observation that the gene expression profiles were highly similar at the two differentiation stages in both studies. Notable, more than half of the most highly expressed genes in both Meyer et al. and Wu et al. studies contained Runx2 peaks at their promoters, emphasizing that Runx2 has an important role in maintaining the phenotype and gene expression profile of osteoblastic cells. Based on our analysis, the minor difference between the studies was that there were very few unique Runx2 peaks in undifferentiated cells in Wu et al. study, while there was a significant number of unique peaks in day 0 sample in Meyer et al. study.

In Meyer's study close to half of the 721 DE transcripts between the differentiation stages overlapped with Runx2 peak-associated genes either at POB, OB or at both states. Similarly to our results, when they looked at the 20% most highly or lowly expressed genes, they found the most highly expressed genes in both POBs and OBs to be significantly more enriched with Runx2 peaks. In the original article of Wu et al. in

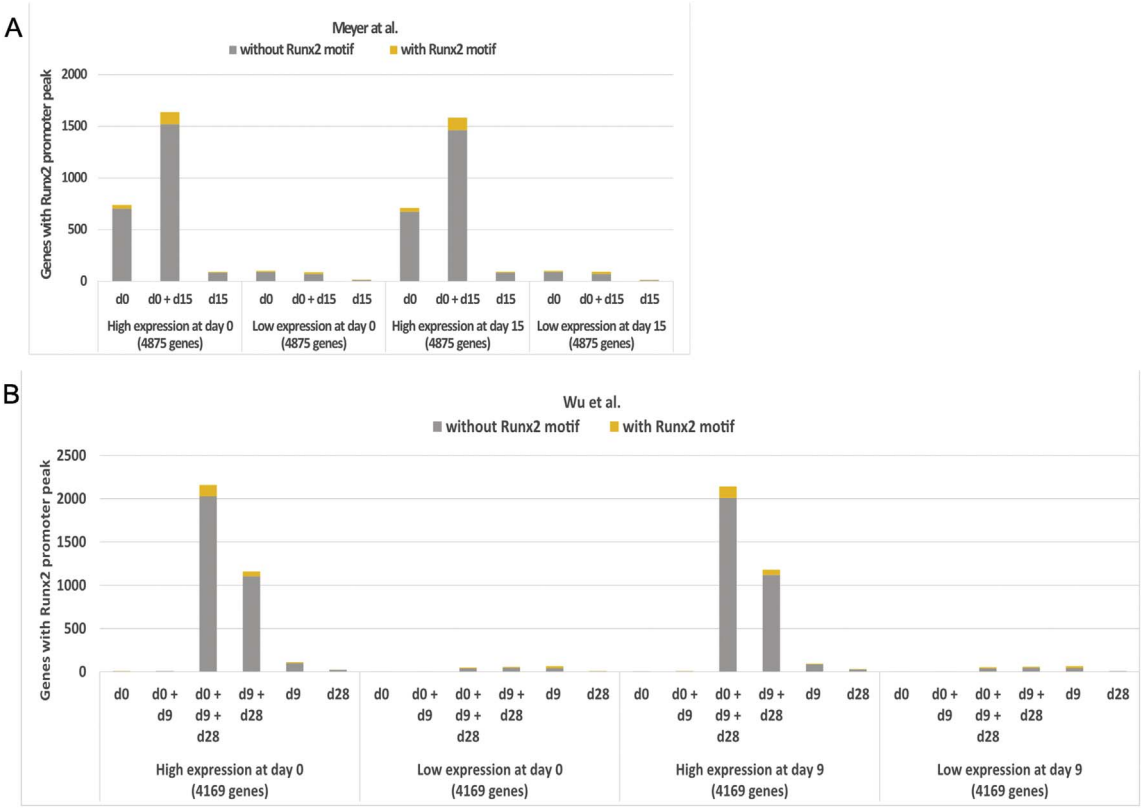

Fig. 6. Distribution of Runx2 peaks on promoters of 20% highest/lowest expressed genes at different time points in mouse MC3T3-E1 cells, A) in Meyer et al. and B) in Wu et al. study. Yellow color indicates the proportion of peaks containing Runx2 motifs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

turn reported 159 Runx2 responsive genes in a Runx2 shRNA knockdown experiment of 9d differentiated cells. Most of these shRunx2 downregulated genes showed higher Runx2 enrichment close to the TSS than the upregulated ones, which is in line with our observation that the promoters of highly expressed genes are more occupied by Runx2 than the promoters of genes with low expression. Despite the abundance of Runx2 at gene promoters of highly expressed genes, the number of shRunx2 responsive genes seems to be surprisingly low, suggesting other than direct transcriptional effects of Runx2 binding at most of the promoters. Accordingly, in SaOS osteosarcoma cells abundant Runx2 occupancy at gene promoters was previously demonstrated not to be directly correlated to gene expression nor highly Runx2 occupied genes to be responsive to Runx2 depletion (van der Deen et al., 2012). Interestingly, in our analysis, a consensus Runx2 binding motif was present only in a minority of the promoter areas occupied by Runx2 (shown also in Supplemental Tables III and IV and discussed in Section 3.2.6), further suggesting collaboration or co-regulation with other proteins at these sites (Fig. 6).

4. Conclusions

We performed a small-scale meta-analysis of three published osteoblast Runx2 ChIP-seq data sets of mouse and human origin and in addition evaluated the similarity of their corresponding gene transcription data produced by DNA microarrays or RNA-seq. From the technical point of view, both ChIP-seq data and gene expression data from two different MC3T3-E1 clones showed high degree of similarity, suggesting that overall ChIP-seq assay even in these difficult samples can produce reliable and reproducible data. Moreover, our data demonstrates relatively small overall variation in the phenotypes of these cells. By using similar data processing and analysis pipeline to all data sets, we could confirm some common features of Runx2 binding reported in the original studies. First, while Runx2 binding is enriched in the genic context including e.g. approximately 20% of all binding sites at gene promoters, relatively high proportion of Runx2 peaks remain in the intergenic genomic locations. Secondly, a substantial portion of Runx2-occupied genomic regions do not contain consensus Runx2 DNA binding motifs, suggesting the Runx2 motif to exhibit even greater variability than previously shown, or that other DNA binding mechanisms and/or collaborative binding with other protein complexes may play an important role in determining Runx2 occupancy at specific sites. Third, majority of the highly expressed genes in osteoblasts contain bound Runx2 at their promoters, however the regulatory role of Runx2 at these sites remain open. In addition, we evaluated interspecies differences in Runx2 binding patterns in mouse and human samples. In general, Runx2 binding features in the human sample resembled closely the observations from mouse data. Importantly, when mouse and human Runx2 binding at the vicinity of genes was compared, approximately 40% of the genes showing highest enrichment of Runx2 were in common, suggesting high degree of similarity and thus conservation of Runx2 regulated mechanisms between the species.

Regarding gene expression profiling, mouse MC3T3-E1 cells showed similar gene expression profiles in two different laboratories, but the differentiation stage caused only minor changes in the global expression profile in these cells. Moreover, human iMSCs and mouse MC3T3-E1 gene expression profiles and enriched pathways showed substantial similarity, suggesting these cell models to represent osteoblastic phenotype in a similar manner. However, there were also some unique cellular processes activated in each sample, highlighting the limitations of in vitro models and importance of studying and reproducing biologically relevant observations in more than one cell line. Finally, our study demonstrates that objective re-analysis and pooling of independent genome-wide NGS datasets is very important to validate the findings of individual studies. With a higher number of replicates and larger pool of data this type of analyses also allow for identification of the true target promoters/genes and pathways and mechanisms of transcriptional regulation that might be otherwise missed.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gene.2017.05.028.

Conflict of interest

The authors declare no conflict of interest.

Funding

This study was funded by the Academy of Finland (R.K.: 298625, 268535, 139165), Emil Aaltonen Foundation, Sigrid Juselius Foundation and the Finnish Medical Foundation.

References

Andrews, S., 2010. FastQC: a quality control tool for high throughput sequence data. (WWW Document) Bioinformatics(doi: http://www.dx.doi.org/citeulike-article-id:11583827).

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Stat Methodol. 57, 289–300.

Cao, Y., Yao, Z., Sarkar, D., Lawrence, M., Sanchez, G.J., Parker, M.H., MacQuarrie, K.L., Davison, J., Morgan, M.T., Ruzzo, W.L., Gentleman, R.C., Tapscott, S.J., 2010. Genome-wide MyoD binding in skeletal muscle cells: a potential for broad cellular reprogramming. Dev. Cell 18, 662–674. http://dx.doi.org/10.1016/j.devcel.2010.02.014.

Cheng, Y., Ma, Z., Kim, B.-H., Wu, W., Cayting, P., Boyle, A.P., Sundaram, V., Xing, X., Dogan, N., Li, J., Euskirchen, G., Lin, S., Lin, Y., Visel, A., Kawli, T., Yang, X., Patacsil, D., Keller, C.A., Giardine, B., Consortium, Mouse E.N.C.O.D.E., Kundaje, A., Wang, T., Pennacchio, L.A., Weng, Z., Hardison, R.C., Snyder, M.P., 2014. Principles of regulatory information conservation between mouse and human. Nature 515, 371–375. http://dx.doi.org/10.1038/nature13985.

Cohen Jr., M.M., 2009. Perspectives on RUNX genes: an update. Am. J. Med. Genet. A 149A, 2629–2646. http://dx.doi.org/10.1002/ajmg.a.33021.

van der Deen, M., Akech, J., Lapointe, D., Gupta, S., Young, D.W., Montecino, M.A., Galindo, M., Lian, J.B., Stein, J.L., Stein, G.S., van Wijnen, A.J., 2012. Genomic promoter occupancy of runt-related transcription factor RUNX2 in osteosarcoma cells identifies genes involved in cell adhesion and motility. J. Biol. Chem. 287, 4503–4517. http://dx.doi.org/10.1074/jbc.M111.287771.

Ducy, P., 2000. Cbfa1: a molecular switch in osteoblast biology. Dev. Dyn. 219, 461–471(doi:http://www.dx.doi.org/2-C).

Håkelien, A.M., Bryne, J.C., Harstad, K.G., Lorenz, S., Paulsen, J., Sun, J., Mikkelsen, T.S., Myklebost, O., Meza-Zepeda, L.A., 2014. The regulatory landscape of osteogenic differentiation. Stem Cells 32, 2780–2793. http://dx.doi.org/10.1002/stem.1759.

Handoko, L., Xu, H., Li, G., Ngan, C.Y., Chew, E., Schnapp, M., Lee, C.W., Ye, C., Ping, J.L., Mulawadi, F., Wong, E., Sheng, J., Zhang, Y., Poh, T., Chan, C.S., Kunarso, G., Shahab, A., Bourque, G., Cacheux-Rataboul, V., Sung, W.K., Ruan, Y., Wei, C.L., 2011. CTCF-mediated functional chromatin interactome in pluripotent cells. Nat. Genet. 43, 630–638. http://dx.doi.org/10.1038/ng.857.

Heinz, S., Glass, C.K., 2012. Roles of lineage-determining transcription factors in establishing open chromatin: lessons from high-throughput studies. Curr. Top. Microbiol. Immunol. 356, 1–15. http://dx.doi.org/10.1007/82_2011_142.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., Glass, C.K., 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell 38, 576–589. http://dx.doi.org/10.1016/j.molcel.2010.05.004.

Hong, J.-H., Hwang, E.S., McManus, M.T., Amsterdam, A., Tian, Y., Kalmukova, R., Mueller, E., Benjamin, T., Spiegelman, B.M., Sharp, P.A., Hopkins, N., Yaffe, M.B., 2005. TAZ, a transcriptional modulator of mesenchymal stem cell differentiation. Science 309, 1074–1078. http://dx.doi.org/10.1126/science.1110955.

Huang, J., Zhao, L., Xing, L., Chen, D., 2009. MicroRNA-204 regulates Runx2 protein expression and mesenchymal progenitor cell differentiation. Stem Cells 28 (N/A-N/). http://dx.doi.org/10.1002/stem.288.

Jonason, J.H., Xiao, G., Zhang, M., Xing, L., Chen, D., 2009. Post-translational regulation of Runx2 in bone and cartilage. J. Dent. Res. 88, 693–703. http://dx.doi.org/10.1177/0022034509341629.

Kim, J.-H., Jang, J.-W., Lee, Y.-S., Lee, J.-W., Chi, X.-Z., Li, Y.-H., Kim, M.-K., Kim, D.-M., Choi, B.-S., Kim, J., Kim, H.-M., van Wijnen, A., Park, I., Bae, S.-C., 2014. RUNX family members are covalently modified and regulated by PIAS1-mediated sumoylation. Oncogene 3, e101. http://dx.doi.org/10.1038/oncsis.2014.15.

Landin-Malt, A., Benhaddou, A., Zider, A., Flagiello, D., 2016. An evolutionary, structural and functional overview of the mammalian TEAD1 and TEAD2 transcription factors. Gene 591, 292–303. http://dx.doi.org/10.1016/j.gene.2016.07.028.

Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B.E., Bickel, P., Brown, J.B., Cayting, P., Chen, Y., DeSalvo, G., Epstein, C., Fisher-Aylor, K.I., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A.J., Hoffman, M.M., Iyer, V.R., Jung, Y.L., Karmakar, S., Kellis, M., Kharchenko, P.V., Li, Q., Liu, T.,

Liu, X.S., Ma, L., Milosavljevic, A., Myers, R.M., Park, P.J., Pazin, M.J., Perry, M.D., Raha, D., Reddy, T.E., Rozowsky, J., Shoresh, N., Sidow, A., Slattery, M., Stamatoyannopoulos, J.A., Tolstorukov, M.Y., White, K.P., Xi, S., Farnham, P.J., Lieb, J.D., Wold, B.J., Snyder, M., 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. http://dx.doi.org/10.1101/gr.136184.111.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359. http://dx.doi.org/10.1038/nmeth.1923.

Lian, J.B., Stein, J.L., Stein, G.S., van Wijnen, A.J., Montecino, M., Javed, A., Gutierrez, S., Shen, J., Zaidi, S.K., Drissi, H., 2003. Runx2/Cbfa1 functions: diverse regulation of gene transcription by chromatin remodeling and co-regulatory protein interactions. Connect. Tissue Res. 44 (Suppl. 1), 141–148.

Lin, Y.C., Jhunjhunwala, S., Benner, C., Heinz, S., Welinder, E., Mansson, R., Sigvardsson, M., Hagman, J., Espinoza, C.A., Dutkowski, J., Ideker, T., Glass, C.K., Murre, C., 2010. A global network of transcription factors, involving E2A, EBF1 and Foxo1, that orchestrates B cell fate. Nat. Immunol. 11, 635–643. http://dx.doi.org/10.1038/ni.1891.

Long, F., 2012. Building strong bones: molecular regulation of the osteoblast lineage. Nat. Rev. Cell Biol. 13, 27.

Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550. http://dx.doi.org/10.1186/s13059-014-0550-8.

Matsumoto, Y., La Rose, J., Kent, O.A., Wagner, M.J., Narimatsu, M., Levy, A.D., Omar, M.H., Tong, J., Krieger, J.R., Riggs, E., Storozhuk, Y., Pasquale, J., Ventura, M., Yeganeh, B., Post, M., Moran, M.F., Grynpas, M.D., Wrana, J.L., Superti-Furga, G., Koleske, A.J., Pendergast, A.M., Rottapel, R., 2016. Reciprocal stabilization of ABL and TAZ regulates osteoblastogenesis through transcription factor RUNX2. J. Clin. Invest. http://dx.doi.org/10.1172/JCI87802.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., Bejerano, G., 2010. GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. 28, 495–501. http://dx.doi.org/10.1038/nbt.1630.

Meyer, M.B., Benkusky, N.A., Lee, C.H., Pike, J.W., 2014a. Genomic determinants of gene regulation by 1,25-dihydroxyvitamin D3 during osteoblast-lineage cell differentiation. J. Biol. Chem. 289, 19539–19554. http://dx.doi.org/10.1074/jbc.M114.578104.

Meyer, M.B., Benkusky, N.A., Pike, J.W., 2014b. The RUNX2 cistrome in osteoblasts: characterization, down-regulation following differentiation, and relationship to gene expression. J. Biol. Chem. 289, 16016–16031. http://dx.doi.org/10.1074/jbc.M114.552216.

Mikkelsen, T.S., Xu, Z., Zhang, X., Wang, L., Gimble, J.M., Lander, E.S., Rosen, E.D., 2010. Comparative epigenomic analysis of murine and human adipogenesis. Cell 143, 156–169. http://dx.doi.org/10.1016/j.cell.2010.09.006.

Nakashima, K., Zhou, X., Kunkel, G., Zhang, Z., Deng, J.M., Behringer, R.R., de Crombrugghe, B., 2002. The novel zinc finger-containing transcription factor osterix is required for osteoblast differentiation and bone formation. Cell 108, 17–29.

Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842. http://dx.doi.org/10.1093/bioinformatics/btq033.

Ramsköld, D., Wang, E.T., Burge, C.B., Sandberg, R., 2009. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. PLoS Comput. Biol. 5, e1000598. http://dx.doi.org/10.1371/journal.pcbi.1000598.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., Smyth, G.K., 2015. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43, e47. http://dx.doi.org/10.1093/nar/gkv007.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., Mesirov, J.P., 2011. Integrative genomics viewer. Nat. Biotechnol. 29, 24–26. http://dx.doi.org/10.1038/nbt.1754.

Rojas, A., Aguilar, R., Henriquez, B., Lian, J.B., Stein, J.L., Stein, G.S., van Wijnen, A.J., van Zundert, B., Allende, M.L., Montecino, M., 2015. Epigenetic control of the bone-master Runx2 gene during osteoblast-lineage commitment by the histone demethylase JARID1B/KDM5B. J. Biol. Chem. 290, 28329–28342. http://dx.doi.org/10.1074/jbc.M115.657825.

Skarn, M., Noordhuis, P., Wang, M.Y., Veuger, M., Kresse, S.H., Egeland, E. V, Micci, F., Namlos, H.M., Hakelien, A.M., Olafsrud, S.M., Lorenz, S., Haraldsen, G., Kvalheim, G., Meza-Zepeda, L.A., Myklebost, O., 2014. Generation and characterization of an immortalized human mesenchymal stromal cell line. Stem Cells Dev. 23, 2377–2389. doi:http://dx.doi.org/10.1089/scd.2013.0599.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P., 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. 102, 15545–15550. http://dx.doi.org/10.1073/pnas.0506580102.

Sudo, H., Kodama, H.A., Amagai, Y., Yamamoto, S., Kasai, S., 1983. In vitro differentiation and calcification in a new clonal osteogenic cell line derived from newborn mouse calvaria. J. Cell Biol. 96, 191–198.

Tang, Y., Feinberg, T., Keller, E.T., Li, X.-Y., Weiss, S.J., 2016. Snail/slug binding interactions with YAP/TAZ control skeletal stem cell self-renewal and differentiation. Nat. Cell Biol. 18, 917–929. http://dx.doi.org/10.1038/ncb3394.

Varelas, X., 2014. The Hippo pathway effectors TAZ and YAP in development, homeostasis and disease. Development 141, 1614–1626. http://dx.doi.org/10.1242/dev.102376.

Wang, D., Christensen, K., Chawla, K., Xiao, G., Krebsbach, P.H., Franceschi, R.T., 1999. Isolation and characterization of MC3T3-E1 preosteoblast subclones with distinct in vitro and in vivo differentiation/mineralization potential. J. Bone Miner. Res. 14, 893–903. http://dx.doi.org/10.1359/jbmr.1999.14.6.893.

Wilson, M.D., Odom, D.T., 2009. Evolution of transcriptional control in mammals. Curr. Opin. Genet. Dev. http://dx.doi.org/10.1016/j.gde.2009.10.003.

Wu, H., Whitfield, T.W., Gordon, J.A., Dobson, J.R., Tai, P.W., van Wijnen, A.J., Stein, J.L., Stein, G.S., Lian, J.B., 2014. Genomic occupancy of Runx2 with global expression profiling identifies a novel dimension to control of osteoblastogenesis. Genome Biol. 15 (R52-2014-15-3-r52). http://dx.doi.org/10.1186/gb-2014-15-3-r52.

Yang, X., Bentik, S., Scheid, S., Spang, R., 2006. Similarities of ordered Gene lists. J. Bioinforma. Comput. Biol. 4, 693–708. http://dx.doi.org/10.1142/S0219720006002120.

Yang, D., Okamura, H., Nakashima, Y., Haneji, T., 2013. Histone demethylase Jmjd3 regulates osteoblast differentiation via transcription factors Runx2 and osterix. J. Biol. Chem. 288, 33530–33541. http://dx.doi.org/10.1074/jbc.M113.497040.

Yu, G., Wang, L.G., He, Q.Y., 2015. ChIP seeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics 31, 2382–2383. http://dx.doi.org/10.1093/bioinformatics/btv145.

Zaidi, S.K., Sullivan, A.J., Medina, R., Ito, Y., van Wijnen, A.J., Stein, J.L., Lian, J.B., Stein, G.S., 2004. Tyrosine phosphorylation controls Runx2-mediated subnuclear targeting of YAP to repress transcription. EMBO J. 23, 790–799. http://dx.doi.org/10.1038/sj.emboj.7600073.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W., Liu, X.S., 2008. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 9, R137. http://dx.doi.org/10.1186/gb-2008-9-9-r137.